Simple versus complex selection rules for forecasting many time series

Robert Fildes[1] and Fotios Petropoulos[1,*]

June, 2014

Abstract

A major problem for many organisational forecasters is to choose the appropriate forecasting method for a large number of time series. Various selection rules have been proposed in order to enhance forecasting accuracy. The simpler approach for model selection involves the identification of a single method, which is applied to all data series in an aggregate manner, without taking into account the specific characteristics of a single series. On the other hand, individual selection includes the identification of the best method for each series, though it is more computationally intensive. Moreover, a simple combination of methods also provides an operational benchmark. The current study explores the circumstances under which individual model selection is beneficial and when this approach should be preferred to aggregate selection or combination. The superiority of each approach is analysed in terms of data characteristics, existence or not of a dominant method and stability of the competing methods' comparative performance. In addition, the size and composition of the pools of methods under consideration are examined. In order to assess the efficacy of individual model selection in the cases considered, simple selection rules are proposed, based on within-sample best fit or best forecasting performance for different forecast horizons. The analysis shows that individual selection works best when specific sub-populations of data are

[1] Lancaster Centre for Forecasting, Department of Management Science, Lancaster University Management School, Lancaster, LA1 4YX, UK
[*] corresponding author: f.petropoulos@lancaster.ac.uk

considered (e.g. trended or seasonal series), but also when the alternative methods' comparative performance is stable over time. A case study demonstrates the efficiency of the recommended selection strategy.

Keywords: automatic model selection, comparative methods, extrapolative methods, combination, stability.

1. Introduction and literature review

Forecasters regularly face the question of choosing from a set of alternative forecasting methods. Where the task the forecaster faces is one of forecasting many series repetitively automatic approaches to selecting the appropriate method are needed – the forecaster has insufficient time to devote to selection for each time series in any one time period. The forecasting methods usually considered are simple, one of a limited range of extrapolative methods including such standbys as exponential smoothing. Two distinct approaches have been proposed for dealing with this problem: *aggregate selection* where the totality of data series are analysed and a method chosen and then applied subsequently to all the time series and *individual selection*, where, for a particular series, each method is compared and the best chosen to produce forecasts for that series (Fildes, 1989). Aggregate selection has the benefit of simplicity but in principle each different time series with its different characteristics (e.g. trend and seasonality, stability) would be better forecast by an individual model that matches those characteristics. Does individual selection generate these expected benefits in terms of improved accuracy? Fildes (2001) shows that if selection could be done perfectly then the gains would be substantial. So the question is worth asking – can practical model selection algorithms, that will lead to forecasting accuracy gains, be implemented? Is the additional effort and added complexity of adopting an individual selection process worthwhile? Additionally, the question is important because simple selection algorithms are implemented in commercial software such as SAP APO-DP.

The task of selecting an appropriate forecasting method is first conditioned by the problem context and the data available. Armstrong (2001), and Ord and Fildes (2013) providing a simplified version, have proposed selection trees that aim to guide the forecaster to an appropriate set of methods. The current study considers the more limited case of choosing between extrapolative forecasting methods where substantial data are available on which to

base the choice. This problem has a long history of research, primarily by statisticians. Broadly, the approach adopted is to assume a particular class of model where selection is to take place within that class, for example within the class of ARIMA models. Accuracy measures based on within-sample fit to the available data are used in the selection, modified in various ways to take into account the number of estimated parameters in each of the models, penalising more complex models. AIC and BIC are two widely used information criteria. Both are based on the likelihood, including a penalty depending on the number of model parameters. As such, AIC and BIC deal with the trade-off between complexity and goodness of fit of the model. Assuming normal errors, minimizing the AIC is asymptotically equivalent to minimizing the one-step-ahead forecast Mean Squared Error.

From the early days of forecasting comparisons, the issue of the strength of the relationship between out-of-sample forecasting accuracy (on the test data) and in-sample fit has been controversial with first Makridakis and Winkler (1989) and then Pant and Starbuck (1990) arguing that little if any relationship exists. Pant and Starbuck examine three different measures of fit and corresponding measures of forecasting performance with mean squared fitted error proving a particularly inadequate guide. But the other measures were not much better. If in-sample fit is inadequate as these authors have argued, then an alternative approach to selection is clearly needed. In response, the forecasting literature became increasingly satisfied with the naïve principle that what has forecast the most accurately, will forecast the most accurately on the out-of-sample data. To operationalize selection based on out-of-sample performance, the available data should be broken into the data used to fit the model (often called the training data), the data used to provide an estimate of out-of-sample fit (the validation data) and the test data where various selection approaches can then be compared.

Beyond the examination of in-sample measures of fit and their link to performance on test data, earlier empirical research has been sparse. One distinct approach has been to use the data characteristics of the series to predict performance with Shah (1997) and Meade (2000) demonstrating some success, but such selection rules are complex. Collopy and Armstrong (1992) also use series characteristics to develop rules that combine various extrapolative models depending on the data conditions. Rule-based forecasting has shown promising performance in various empirical comparisons. A contrasting approach which benefits from simplicity is to consider past performance as the critical factor predicting future performance. A recent contribution is from Billah, King, Snyder and Koehler (2006) who consider selection within the class of exponential smoothing models where an overarching general model exists. Their results for a subset of the M3 data demonstrated that information criteria outperform the use of the validation data in selection: but as they remark, the sample of out-of-sample validation forecasts is small, which, they conjecture, might explain their findings. The differences are also small between selection methods so a reasonable conclusion to draw might be that selection is not worthwhile – but that only applies to their particular data set and the extrapolative forecasting methods they considered. However, with the M3 data, automatic individual selection based on Forecast Pro's algorithm (Goodrich, 2000) had earlier proved effective, beating most aggregate selection approaches post hoc. Further work has been reported by Crone and Kourentzes (2011) who, using a different data set, demonstrate the benefits of using out-of-sample error measures compared with in-sample. In short, earlier research has produced conflicting results.

The contradictory findings leads to the following observations: individual selection can never be worthwhile if a dominant aggregate forecasting method is identified in the data set. Moreover, selecting the best method individually will not provide significant benefits if the methods under consideration produce similar forecasts.

This paper aims to provide evidence on the effectiveness of the various selection criteria introduced. Following on from the above argument, there is the need to vary the methods considered for selection and also the data sets on which selection algorithms are tested. In section 2 the forecasting methods considered in the selection comparisons and the error measures being used to assess their accuracy are introduced. Section 3 considers the meta-data set (part of the M3 database), introduces the simple selection rules and also explains the rationale behind the different segments examined. Section 4 contains the empirical results and provides a discussion of the results. Section 5 comments on the practical implications and limitations of the current research, including a brief case study. The conclusions are drawn out in section 6. The key question to be addressed is under which circumstances can individual selection rules generate accuracy benefits.

2. Forecasting methods and accuracy metrics

2.1 Extrapolative forecasting methods

In this evaluation of selection methods typical practice is emulated such as that embedded in forecasting software. The forecasting methods considered are therefore chosen broadly to represent standard approaches but are not themselves nested in an overall model, such as in the exponential smoothing class of Billah et al. (2006). They have been chosen from those considered in the forecasting competitions, in particular the M3 competition (Makridakis and Hibon, 2000) in which larger numbers of series have been analysed and a large number of extrapolation methods have been compared. All are practical alternatives in commercial applications. Computer intensive methods such as neural networks have been excluded. Therefore, the focus is on simple extrapolation methods, methods widely used in practice, and also including some that have demonstrated significant performance in past forecasting

exercises. The simplest forecasting technique, Random Walk or Naïve, where the forecast is the latest observation, is therefore included along with widely used models from the exponential smoothing family (ETS, Hyndman, Koehler, Snyder and Grose, 2002), namely Simple Exponential Smoothing (SES), Holt, Holt-Winters, Damped Trend and Damped with multiplicative seasonality. Moreover, despite their limited use in practice, ARIMA models have been included as they remain a standard statistical benchmark.

The exponential smoothing methods are estimated using the *forecast* package for *R* statistical software (Hyndman and Khandakar, 2008). The Automatic ARIMA function (*auto.arima*) implemented in the same package is used to identify and estimate the ARIMA models. The *auto.arima* function conducts a stepwise selection over possible models and returns the best ARIMA model. One could argue that this advantages ARIMA over other methods (such as SES or Holt), as the Automatic ARIMA function already aims to choose the best model from within a broad class of models. Hyndman and Athanasopoulos (2012) describe how the *auto.arima* function works, with details on the options allowed in the ARIMA modelling.

In all cases mentioned above, the methods are applied directly to the raw data. However, in previous large forecasting exercises, such as the M3-Competition (Makridakis and Hibon, 2000), the non-seasonal methods were applied to the deseasonalized data. Deseasonalisation of the data is usually conducted with multiplicative classical decomposition, where the seasonal indices calculated are used for the reseasonalization of the final forecasts. More details on how the deseasonalization is applied in the current research can be found in the Appendix. In order to be in line with the results of this research, simple and widely used models (Naive, SES, Holt and Damped) applied to the seasonally adjusted data, instead of the raw data, are considered. Lastly, the Theta model (Assimakopoulos and Nikolopoulos, 2000),

which was the top performer in M3-Competition, is considered. More details on the Theta model can be found in the Appendix.

The full set of methods considered in this paper, along with the respective short names, is presented in Table 1.

Table 1 here.

2.2 Measuring Forecast Error

Measurement of each method's forecasting performance is needed in two distinct phases of this research. Firstly, the forecasting performance of each method can be calculated over the validation data set (which is defined rigorously in Section 3.1), and these measures can then be used in the selection of an appropriate method. This can be achieved by calculating the past forecasting performance (*PFP*) over the available validation data and across single or multiple lead times (full definitions are provided in the Appendix). The fit of the models in-sample can also be calculated. Secondly, metrics for measuring the performance of the methods and the selection rules are necessary in order to assess the efficacy of the latter. The mean out-of-sample performance is averaged over forecast origins (and potentially over forecast horizons). The general formulae are given in the Appendix.

A wide range of different error measures are available. A summary of the arguments surrounding their differences has recently been given by Davydenko and Fildes (2013). The majority of results reported in this paper are based on Median Absolute Percentage Error (*MdAPE*), where the arithmetic mean of the absolute percentage errors across horizons and origins is calculated for each series, and the median value over all series is selected.

However, the validity of the results is confirmed based on two more measures, Mean Absolute Percentage Error (*MAPE*) and a relative error measure. *MAPE* is averaged over all

series, forecast horizons, and all available forecast origins. *MAPE* has been included here as the most widely adopted in practice (Fildes and Goodwin, 2007). Relative error measures have the advantage of negating the effects of outliers somewhat showing how a forecasting method compares to a benchmark (such as the random walk). Summarizing relative errors across series using geometric mean has proved to be robust and more normally distributed than alternative measures, while also being easily interpretable as showing the average percentage improvement (as measured by the MAE) from using one method compared to the benchmark method. In the current study, the *AvgRelMAE,* as defined by Davydenko and Fildes (2013), is used.

3. Experimental design

3.1 Forecasting procedure and database

Let T denotes the number of observations of an individual time series. Each series considered in this paper is divided in three time intervals. The first interval contains all observation from origin 1 to origin T_1, having a length of T_1, and acts as an initialisation interval, that is the training data. Observations from origins T_1+1 to T_2 are included in the second interval, the validation data, while the third interval contains observations between origins T_2+1 to T. The second and the third intervals have respectively length (T_2- T_1) and (T-T_2). Both corresponding sets of data are used as hold-out samples, meaning that forecasts are produced without prior knowledge of these values. Once the first set of forecasts is produced, using just the T_1 in-sample observations, one additional observation, the first observation of the validation data, is added to the in-sample data, the estimated models updated and new forecasts are calculated. This procedure is repeated until every single observation of the validation and test intervals is embodied into the in-sample vector. In other words, rolling

forecasting is employed, where the forecasts (and selected models) are updated at every single origin. As a result, $(T_2 - T_1) + (T - T_2) = T - T_1$ sets of forecasts are calculated, each one containing h point forecasts, where h denotes the forecasting horizon considered.

The second interval is used only as validation data, in terms of evaluating single extrapolation methods and selecting the most appropriate one for forecasting each series (individual model selection) or a method to apply to all series (aggregate model selection). The third interval is used as both test data for the final evaluation of the selection rules proposed later in this paper, and, as the forecast origin is rolled forward, the associated validation data set is also extended. Multiple lead-time forecasting enables the set-up of simple selection rules that apply to the various forecasting horizons.

The data series selected for this study are a sub-set of the monthly M3-Competition data set, where the total length of available observations is equal to or greater than T=126, giving a total of 998 series. Data series longer than the desired 126 observations are truncated. T_1 is set to 48 and T_2 to 90. Thus, the first set of forecasts is calculated from time origin T_1 (=48). The forecasting horizon was set to h=18 periods ahead, to correspond with earlier analyses of the same data (Makridakis and Hibon, 2000).

Upon the calculation of the point forecasts for each method using the first 48 data points, an additional point is added and a new set of point forecasts are calculated for each approach. This procedure is repeated until the origin 108, where the last 1-18 steps-ahead forecasts are produced. The remaining data points (observations 109 to 126) are only used in the evaluation of the last origin's forecasts. So, in total, 18 point forecasts are produced for each origin (origins 48-108, 61 origins in total) and for each method (12), while the out-of-sample performance of all methods plus the model selection rules are evaluated through observations T_2+1 to T, the test data. The selection of the most appropriate method, based on past forecast

performance of the methods in hand as calibrated over the validation data set, takes place at observations T_1+1 (=49) through T_2+k (periods 90 to 108 with k =0 to 18 indexing the test data). These 998 series provide the meta data set within which subsets of the data will be examined.

3.2 Choosing a best method

The objective of any selection rule is to choose the method at time t with the most promising performance. For the purposes of the current research, various simple selection rules are considered, based on the past forecasting performance (*PFP*) of each method (for different lead times). Assuming that model selection will be performed at the T_2+k origin, the *PFP* is measured between origins T_1 to T_2+k as is the fitted performance. The method to be selected is the one with the most promising past performance. The four simple rules implemented and examined in this research are defined as follows:

Rule 1. Use the method with best fit as measured by the minimum one-step ahead in-sample Mean Squared Error (using all the data up to the forecast origin).

Rule 2. Use the method with the best out-of-sample 1-step-ahead forecast error, in terms of Mean Absolute Percentage Error, and apply that method to forecast for all lead times.

Rule 3. Use the method with best out-of-sample h-step-ahead forecast, in terms of Mean Absolute Percentage Error, and apply this method to forecast for the same lead time.

Rule 4. Use the best out-of-sample method to forecast for all lead times as measured by Mean Absolute Percentage Error averaged over forecast horizons, 1 through h.

While the mathematical expressions are complex (as shown in the appendix) the ideas behind them are simple. Rule 1 selects the method that has fitted the data best (in-sample)

and applies this method to forecasting from forecast origin *t* over the next forecast horizons. Rules 2 to 4 select the best method depending on how the methods performed as measured on past forecast performance over the validation data set up to the forecast origin. Rules 1, 2 and 4 ignore any horizon effects while only Rule 3 attempts to match selection to the forecast horizon. The selections derived from these rules are updated over the test data set (i.e. as *k* increases), including all available errors from origins T_1 to T_2+k-1. Moreover, the proposed rules can be applied to aggregate selection where the error measures are summarized over all series and the best method is applied to all series, or to individual selection where a particular method is chosen for each series.

3.3 Research questions and preliminary analysis

The main objective of the current research is to investigate the conditions under which model selection may be beneficial. In order to achieve this objective, three primary segmentations of the available time series are considered. Firstly, data are classified as trended or not trended and seasonal or not seasonal. These categorisations have been chosen a priori based on the fact that some of the models are designed to incorporate trend and seasonal etc (e.g. ARIMA, Holt-Winters) whilst others (e.g. Random walk, Simple Exponential Smoothing) will introduce unnecessary error when applied to series with these characteristics. A further feature, believed to affect relative performance, is the predictability of the time series. A specific time series is defined as unpredictable if the performance of the non-seasonal Random Walk forecasting method (method 1) is better than the median performance of all other methods under investigation as defined by Mean Absolute Error in the validation data (from origins T_1 to T_2) for all forecasting horizons. Note that this classification is available to us ex ante and does not use the test data.

In terms of trend, the robust Cox-Stuart test is performed on the 12-period centred moving average, to remove any contamination from seasonality. Lastly, the potential seasonal behaviour of the monthly series considered is tested by Friedman's non-parametric test. As a result, six segments of the time series data set are considered, namely "predictable", "unpredictable", "trended", "non-trended", "seasonal" and "non-seasonal" and this suggests the first research question.

RQ1. *Is individual model selection more effective when applied to groups of time series with specific characteristics?*

A second factor that may limit the value of individual selection is the number of models included in the pool of alternatives. Effectively a variant of over-fitting, the more models included, the higher the probability that the wrong model is chosen due to the randomness in the data. Given that the largest pool can be structured with all methods introduced in Section 2.1 (twelve in total), every possible combination of smaller pools of two (2) up to twelve (12) methods is also examined. For example, in the case of a pool of methods equal to four (4), all 495 possible pools of methods are checked, the number of 4-combination in a set of 12 or $\binom{12}{4}$. This leads to the second research question:

RQ2. *What are the effects on individual selection of including more methods in the pool under consideration?*

Many of the methods included in typical extrapolative selection competitions produce similar forecasts which may be difficult to distinguish using a selection rule. An analysis of the correlation of errors produced by the methods revealed that many methods are highly correlated, most obviously those with similar seasonality components. Holt and Holt Winters have the fewest high correlations with the remaining methods. Selection between methods

that produce similar forecasts cannot prove valuable. On the other hand, when selecting among methods that produce uncorrelated forecasting errors, rules have better chance of discriminating to deliver the best result. Including more uncorrelated options in the selection and letting the rules decide based on the past forecasting performance of each option is intuitively appealing. In that sense, selecting among methods with low to medium correlated outputs and similar levels of accuracy (e.g. DampMult and ARIMA) is more promising. The average error correlation from the various methods participating in a specific pool is therefore examined. In order to measure the effect of similarity between the methods included in a specific combination, the combinations in each pool size are separated into high and low correlated; a certain combination is considered as highly correlated if the average correlation of the methods' outputs is equal to or greater than 0.7. Thus, the following research question deals with the effect of correlation among methods.

RQ3. *Do pools of methods with low correlation, in terms of forecast error, provide better forecasting performance when individual selection rules are considered compared to more highly correlated pools?*

Individual selection would be unlikely to be beneficial when a single method is dominant for the obvious reason that if a single method was appropriate for all series, selection rules would be dominated by the effects of noise. A second segmentation is, therefore, considered: to divide the data series into two groups in terms of the performance of one of the best methods. The aim here is identify sub-populations where a dominant method exists (or not). For this purpose, the Theta method is chosen. In the M3 Competition Theta had the best performance over the 1,428 monthly series and also performs well over the subset of 998 series. The threshold for a specific time series to be grouped in one of the two groups will be the Theta model's achievement to be ranked (or not) among the top three (out of twelve)

methods. In other words, past forecasting performance, as measured by Mean Absolute Error for the validation data, must be lower (or higher) than the value of the first quartile, that is:

1st Group (Dominant method): Theta's performance in the top three (measured by MAE)

2nd Group (Non-dominant method): Theta's performance outside the top three

Note that the selection of Theta method as the base method for segmenting the data is data set dependent.

RQ4. *Individual method selection is of most value when no dominant method is identified across the population.*

The performance of the different methods is also analysed for their stability. Using the validation data, the error can be calculated for each data point. Then, stability in a specific series can be measured by the average (across time origins) Spearman's rank correlation coefficient where the ranked performance of methods at each forecast origin is correlated to the rank of the average performance of the method summarized across all origins. A value of 1 implies the rankings of all methods remain the same over time. The median of the stability measure is 0.45 with range 0.01 to 0.91. Thus, the 998 series may be further segmented in regards to the stability of methods' performance. A series is defined as stable when its Spearman's rho falls in the top quartile of the data set. As a result, a final research question is suggested:

RQ5. *Individual selection is only effective compared to aggregate selection when relative performance in the pool of methods under consideration is stable.*

In total, segmentations of data considered in the current research are summarized in Table 2, where the respective populations of the groups are displayed.

Table 2 here.

4. Empirical results

4.1 Out-of-sample performance of methods

Firstly, the out-of-sample performance of the forecasting methods is examined using error measures averaged across all origins (T_2 to T_2+18) and lead times (1 to 18). This is achieved by using a rolling origin design with forecasts made for each point in the validation and test data sets.. Table 3 presents the results when *MdAPE* is selected as error measure. Each row refers to a single extrapolation method (please, refer to Table 1 for the abbreviations). At the same time, each column refers to a specific segmentation of the data, as described in Table 2.

Even a quick view of this table unveils some very interesting observations. Firstly, across all segments, the best performance, in terms of accuracy is recorded for Theta followed by SES, when applied on the seasonally adjusted data, and seasonal versions of Damped. Theta and Deseasonalized exponential smoothing, correlated at 0.98, perform very similarly for all segments. Over all series, Holt and Naive demonstrate the worst performance, neglecting as they do seasonality and, for the case of Naive, trend. The largest differences across the methods are recorded for predictable and seasonal series, where methods with specific features, such as the ability to handle seasonality, perform much better than benchmark methods. On the other hand, simpler methods catch up with more complex ones when unpredictable or non-seasonal series are considered. The presence of trend or seasonality naturally favour methods with the ability to capture these features, when they persist. SES on deseasonalized data (DExpsmoo) performs well and better than ARIMA.

The segment of data series containing the non-trended series suffers from relatively high levels of inaccuracy (an average *MdAPE* across methods of 16.9% compared with 8.3% overall). As expected, Holt performs second worst (following Naive), failing to estimate the zero trend. When segmented on the stability in the methods' relative performance, non-seasonal methods performed uniformly poorly suggesting stability in performance is related to the ability of a method to capture persistent seasonality. For the 749 unstable series differences in performance are much smaller.

Finally, the last row presents *MdAPE* values for perfect information, meaning that the best method is selected for each series (individual selection) in an ex-post manner after the comparative accuracy results are known. Possible margins of improvements (perfect information) are between 25 to 30% for all segments, compared to the best method in each segment applied to all series (ex-post aggregate selection). Therefore, individual model selection is worth investigating, as Fildes (2001) had previously argued. In addition, segmentation of the series emphasizes the importance of trend, seasonality and stability.

The out-of-sample performance analysis was performed for two more error measures, *MAPE* and *AvgRelMAE*. The use of *MAPE* results in significantly higher errors, as expected. Increases across the different methods on each segment are consistent, resulting in stable ratios of *MAPE*/*MdAPE*. *AvgRelMAE* confirms the superiority of Theta and DExpsmoo across all series and for most of the segments, with Naive being among the best methods for unpredictable and non-seasonal series (recall that the unpredictable series are defined ex ante where naïve performs well).

Table 3 here.

4.2 Performance of selection rules

In this section, the performance of the various selection rules are presented. The empirical results focus on the number of cases improved by performing individual selection versus two simple benchmarks:

(i) Aggregate selection; this uses the single best method per segment based on the one-step-ahead out-of-sample performance on the validation sample.
(ii) Combination of methods using equal weights.

Thus, the accuracy gains (or losses) by using simple individual selection rules are examined through the percentage of cases where individual selection rules performed better than the above benchmarks. Accuracy is measured through the percentage of cases where forecasting accuracy is improved by individual selection as measured by *MdAPE*. Improvements in more than 50% of the cases are presented with bold typeface. The results are segmented by the sizes of the pools of methods under consideration and by the correlation of methods in a specific pool (e.g. ARIMA and Expsmoo have a low correlation).

Table 4 presents the percentage of cases improved in terms of accuracy when all series are considered. Recall that Rule 1 uses in-sample one-step-ahead fit, Rule 2, one period ahead past forecast performance, Rule 3 matches the lead time in individual selection while Rule 4 takes a more aggregate approach with selection based on average performance over all lead times. So, for example, of the 611 cases of selection using between 2 and 4 low correlated methods, individual selection using Rule 4 was more accurate than aggregate selection (using Rule 4) in 88.1% of these case comparisons and in 90% of cases when compared to simple combinations of the corresponding methods.

The first observation is that the relative number of cases improved by individual selection is higher when the rules are applied to low correlated or uncorrelated methods, especially when small pools are considered. Also as the pools' size increases, individual selection

generally becomes better. At the same time, these improvements are only achieved for Rules 2 to 4, with Rule 1 (best in-sample fit) having the worst performance. Moreover, individual selection always outperforms both aggregate selection and combination in more than 80% of the cases when Rule 4 is applied to methods identified as low correlated.

Table 5 shows the percentage of cases that for each segment individual selection performs better when compared to aggregate selection or combination. For predictable series, aggregate selection works generally better than individual selection, with combination being efficient for medium pools of high correlated methods. Recall the definition of a predictable series is made on the validation data and therefore offers guidance on whether to use individual selection. In addition, for predictable series, individual selection works better than combination only low correlated methods, while higher improvements are observed for larger pools of methods, since in most combinations consistently poorer methods are included. On the other hand, individual selection (with Rule 4) is the best option for unpredictable series.

Table 4 here.

When trended data are examined, individual selection (compared to aggregate selection) seems to work reliably only for Rule 4. Rules 2, 3 and 4 result in significant improvements when contrasting individual selection to a simple combination of methods. One plausible explanation is the ability of selection to identify methods that include trend. As expected, excluding non-trended methods when extrapolating trended series results in better forecasts for simple combinations, though individual selection is still best for smaller pools of methods. On the other hand, when non-trended data are considered, improvements from individual selection in more than 50% of the cases are limited to only low correlated pools of methods.

Improvements for seasonal data are substantial, especially when Rules 3 and 4 are applied, suggesting reliance on 1-step ahead forecasts for seasonal forecasting is unwise. Essentially, selection is capturing the persistent seasonality in the series. Improvements are higher against aggregate selection in the case of highly correlated pools of methods, while, as expected, the reverse is true against the combination of methods. Individual selection does not usually perform well when non-seasonal series are examined.

The results are also analysed by segmenting the series using the performance of a dominant method. The segment containing the series that Theta was in the top three performers is first examined. Unsurprisingly aggregate selection is the best option (although recall Theta is not necessarily included in each case). The advantages of using individual selection against aggregate selection are more apparent when applied to the 'non-dominant method' segment. At the same time, individual selection is a good choice against combination for both 'dominant method' and 'non-dominant method' segments, especially in the case of low correlated methods.

Table 5 here.

Finally, the series are segmented with regards to the stability of methods' ranked performance. With stability in a method's performance it is of course easier to identify the best individual selection. When series with unstable methods' performance are considered, combination typically outperforms selection. Individual selection does, however, improve over aggregate selection especially for low correlated methods, while Rule 4 performs marginally better against combination for high correlated methods.

The same analysis was performed for the two other measures considered (*MAPE* and *AvgRelMAE*). Insights generally hold for most of the segments, with individual selection being preferable over aggregation or combination under the same conditions; however,

differences in the percentage of cases improved are noticeable. Moreover, in some segments (for example, entire data set, predictable and seasonal segments) the number of cases where individual selection is better than aggregate selection or combination decreases, an effect arising mostly when larger pools of methods considered.

So far, the analysis of the results focused on the percentage of cases improved when individual selection is preferred over aggregate selection or combination. However, from a practical viewpoint, it is very important to identify the best approach in terms of absolute performance for each segment. Table 6 presents the best practices for each segment, in terms of the most appropriate approach (individual selection, aggregate selection or combination). Thus, for the entire data set it would be best to apply individual selection (using rule 4) based on a pool of large number of low correlated methods. The specific recommendations are based on the minimization of the *MdAPE* across all possible strategies (selection rules, number of methods and inter-method correlation) within the given segment

Table 6 here.

The relative efficiency of the best practices for each segment is further examined in Table 7. The performance of the best practice is contrasted with the performance of each approach (individual selection with Rule 4, aggregate selection or combination). In addition, the performance of Damped (a suitable benchmark from the various competitions, Fildes, Hibon, Makridakis and Meade, 1998) applied on the deseasonalized data is presented, along with the percentage improvement against this benchmark, if the best practice for each segment were to be applied. In all cases, the performance of alternatives is measured by *MdAPE*. Results indicate that, even if individual selection is the most promising approach for most of the cases, in some segments aggregate selection or combination should be adopted. However, the forecast error of individual selection has in all cases the smallest interquartile ranges,

rendering the forecasts more robust. The suitable selection of the best practice offers improvements over all single-pronged approaches in most of the cases. For example, focusing on the unpredictable segment, if individual selection (Rule 4) is applied, as recommended in Table 6, to a high number of high correlated methods then the resulting *MdAPE* is 8.5%, improved by 11.5%, 3.4% and 2.8% compared to the median performance of aggregate selection, combination and individual selection (Rule 4) respectively. Moreover, best practices for each case, as proposed in Table 6, can lead to significant accuracy improvements against the benchmark (11% for unpredictable series, 5.3% for stable series, and 4.7% for the entire data set).

Table 7 here.

4.3 Discussion

The empirical findings of this study provide some interesting evidence on the efficiency of selection rules. First, segmenting the series helps us to identify suitable sub-populations of data with specific characteristics, where the application of individual selection is more effective (RQ1) compared to the simplest rules of aggregate selection or combination. Individual selection is particularly effective against the two benchmarks for seasonal, trending and 'non-dominant method' series. Individual selection with Rules 1 to 3 does not work well against combination for unstable segments, where the risk averaging aspect of combinations proves, as expected, to work well. Also, its performance is limited when highly correlated methods are used for predictable and non-trended series. Aggregate selection works more effectively than individual approach where a dominant stable method exists, as Fildes (1989) shows in analysing a method, robust trend, designed for the specific non-seasonal trending data set.

RQ2 questioned the effects of including more methods in the pool under consideration. In most cases, when *MdAPE* is considered, improvements for individual selection are recorded when more methods are considered in the selection pool. This was especially evident against combination, because of the incorporation of poorly performing methods, while the selection pool benefits are not disadvantaged by the inclusion of poorer methods.

Low correlated methods offer the foundation of more efficient individual selection compared to aggregate selection or combination. Specifically, when individual selection is contrasted against combination, the number of cases improved from selecting pools containing methods identified as low correlated reaches double the score of the respective highly correlated pools (e.g., Table 5, predictable, non-trended or stable series, Rules 3 and 4). In answer to RQ3 therefore, pools of methods with low correlation generally provide a better foundation for individual model selection. However, in some cases differences are, on average, small (unpredictable and trended series). This leads to the obvious conjecture about whether methods that exclude trend or seasonal should be included in selection schemes for the corresponding segments (i.e. trend or seasonal). Their inclusion does have the advantage of being 'conservative' (Armstrong, Green and Graefe, this issue). Nevertheless, it should be noted that while low correlated pools offer the grounds for individual model selection to improve over aggregate selection or combination, this does not exclude high correlated pools from having the lowest *MdAPE* within a segment (for example, Table 6, stable series).

Aggregate selection is expected to produce better results than individual selection, when a single method displays dominant performance across a specific sample of series (RQ4). The hypotheses is supported through segmenting the data into series where the dominant method achieved (or not) a ranking among the top three methods (out of twelve in total): aggregate selection is better than individual selection in 66% of the cases and is the recommended practice for this segment. The exact opposite is true for the 'non-dominant method' series,

while individual selection outperforming aggregate selection in 74% of the cases. In addition, individual selection displays significant gains over combination (whether a specific method is dominant or not).

Lastly, as expected, when stability in methods' ranked performance is used as a basis of segmentation, individual selection produces more accurate forecasts especially for Rule 4 (RQ5). Stability in performance of methods enables the accurate selection of the most appropriate method individually, with performance improvements of 1% and 8.7% against aggregate selection and combination respectively. This is a direct result from the great differences in the performance of methods over these series (Table 3). On the other hand, for the unstable segment, the combination of methods is the most robust choice, while individual selection with Rule 4 proves to be equally effective when applied to pools of high correlated methods.

In the introduction, no assumption was made as to the effectiveness of the different selection rules, merely noting the existing evidence was conflicting. Empirical results suggest that Rule 1 based on a measure of fit is uniformly ineffective compared to rules based on the validation sample and in particular Rule 4 which looks at aggregate performance averaged over lead times.

5. Case study, practical implications and limitations

5.1 Case Study

In order to demonstrate the practical use of the proposed rules, a case study is presented. Empirical data from a UK based company providing private label household and personal care products developing are examined. In total 251 monthly time series are considered, each one containing a 40-month history. T_1 is set to 24, T_2 is set to 36 and forecasting horizon is

equal to three months. So, the out-of-sample accuracy of the best practices is measured over two sets of forecasts (origins 36 and 37) in a rolling manner, while the past forecasting performance of the methods previously considered is calculated over 12 rolling origins for the first origin, increasing to 13 origins for the second origin.

Table 8 here.

Table 8 presents the performance of the best practice, as defined in Table 6. When the generally low correlation of methods did not allow for selection of a best practice due to the absence of any (high correlated methods) cases, then the appropriate selection approaches for low correlated methods were considered. This performance is compared against the performance of Damped method applied on the deseasonalized data (acting as the benchmark). Across all series, where individual selection (Rule 4) is applied on a high number of low correlated methods, an improvement of 5% is recorded. Improvements are also identified for half of the segments, for example unpredictable series (10.8%), non-trended series (18%), and dominant method (5.8%). However, some segments (most predominantly the seasonal series) result in inferior forecasting performance compared to the benchmark. This is due to three reasons: a) the limited historical information which renders the identification of seasonality difficult, b) the limited number of series featuring in seasonal and stable segments of series, and c) the overall limited forecasting performance of the aggregate selection approach, in contrast to the performance of individual selection and combination on this specific data set. In this case study combination performed very well against both aggregate and individual selection.

5.2 Practical implications and limitations

The insights provided to the use of extrapolative methods can be directly applied to widely used ERPs and Forecasting Support Systems (e.g. SAP), as to further enhance the integrated

automatic selection procedures they have in place. Table 6 presents the appropriate protocol that should be followed, based on the analysis of the long monthly series coming from the M3-Competition. Even if in half of the cases simple and easy to apply approaches should be followed (selection of the appropriate method across all series or the combination of a set of methods), five cases indicate the use of a more complicated selection scheme, where the most appropriate method should be identified for each series separately.

Rules 2, 3 and 4 of individual selection specify a rolling forecasting evaluation for the selection of the optimum method. Coupled with the fact that some real life applications include the extrapolation of many thousands of series, this means that applying individual selection rules could be very computationally intensive. The problem is exacerbated when the frequency of updating forecasts is very high (e.g. daily data). Therefore, the adoption of a specialized system design is important. Such a system would allow the storage of the past forecasting performances of each method and each origin in the purpose-designed database, so that they can be recalled efficiently.

Other drawbacks, when applying these rules in practice, would be the limited data history available. Also, series coming from a specific industry would appear more homogenous in nature and therefore the conclusions of this research might not apply. In order to overcome these problems, the proposed experimental design would need adjustment. In addition, segmentations should be carried out where appropriate. For example, the 48 observations used for initialization purposes could be significantly lessened, as demonstrated in the case study, if the available historical information is limited. Moreover, past forecasting performance could be checked in a smaller rolling window. With regard to defining useful segments, the method that would serve as the cut-off for segmenting the data into ‘dominant’ and ‘non-dominant’ method series should be defined appropriately, based on the rolling out-

of sample performance of all methods. However, limited historical information is always a problem when seasonal data are examined.

6. Conclusions

When forecasting a population of time series, individual selection of the most appropriate method is intuitively appealing and may result in substantial gains. In the current research the circumstances under which selection of an individualized method per series should be preferred to selecting a single method (aggregate selection) for the whole population of series or by a combination of methods were analysed. To explore the conditions when individual selection is most likely to be of benefit, the entire data set was segmented into sub-populations with regard to basic series characteristics (predictability, trend and seasonality). Moreover, the efficacy of individual selection was examined in the case that a specific method is dominant or when the methods' performance are stable across forecast origins. Lastly, the effect of the number of methods taking part in selection (pool size) and the correlation between methods was considered. Based on the above, a protocol for selecting the best possible rule for each segment was proposed and a simplified version evaluated through a case study.

Empirical results, based on the long monthly series of the M3-Competition provided insights with regards to the effectiveness of individual selection versus the simple rules of aggregate selection or combination. When a population of series is divided in sub-populations with specific characteristics, then selection per series is more effective, especially for series identified as seasonal, unpredictable or trended. In addition, individual selection is superior when methods' ranked performance in each series is stable. On the other hand, aggregate selection is the best choice when one single method is dominant across a sub-population of

series, while combination is efficient for predictable and non-trended series when high correlated methods are included in the method set. Finally, with some exceptions, individual selection works better (compared to aggregate selection and combination) when pools of low correlated methods are available.

Of the various individual selection rules considered, Rule 4, which relied on aggregated forecast performance over horizons, proved better than relying on 1-step ahead rules, or even Rule 3 which matched selection to the corresponding horizon. Simply relying on past in-sample performance over the fitted data proved inadequate. This analysis has allowed various practical implications and limitations to be drawn, while a case study on real company data set supported the efficiency of the proposed protocol.

A natural path for future research is to extend the range of methods to include ones with distinctive performance characteristics, such as Neural Networks. Moreover, the selection rules used in this study are only based on model fit and past forecast performance of methods across single or multiple lead times. These could be enhanced by a large number of variables proposed in the literature to characterize a time series (Shah, 1997; Meade, 2000; Adya, Collopy, Armstrong and Kennedy, 2001). Lastly, the current research does not fully take into account the specific features of each extrapolative method, with all pools of possible methods being handled in the same manner. This is done in an attempt to gain a holistic view on the effectiveness of the individual selection rules over the aggregate selection and simple combination of methods (an approach adopted also in commercial software). However, in an operational set up, the inclusion in the selection pool of only the appropriate methods that will match the characteristics of a specific sub-population of data (e.g. trended or seasonal series) seems important. Although an obvious point to a statistician, this principle is not embedded in commercial selection software. Lastly, future work is needed to theoretically

explore why in some segments high correlated pools of methods led to the minimum *MdAPE* and, thus, being the recommended “best practice”.

For many applications selection rules are likely to deliver improved forecast accuracy. While for most sub-populations the gains are not usually large, the reliability is improved. While aggregate selection, perhaps the standard simple rule in application, can clearly deliver where a specific structure characterizes the time series population (e.g. the telecoms data of Fildes (1989)), where the data are more heterogeneous as here in both the M3 and company data sets, individual selection is needed. A final note, the simple rule of combining proved ineffective for most of the segmented data sets, but proved its worth in the case study.

References

Adya, M.; Collopy, F.; Armstrong, J. and Kennedy, M. (2001), 'Automatic identification of time series features for rule-based forecasting', *International Journal of Forecasting*, *17*, 143-157.

Armstrong, J. S., ed. (2001), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Boston and Dordrecht: Kluwer.

Armstrong, J. S.; Green, K. and Graefe, A. 'Golden Rule of Forecasting: Be Conservative', *Journal of Business Research*, this issue.

Assimakopoulos, V. and Nikolopoulos, K. (2000), 'The Theta model: a decomposition approach to forecasting', *International Journal of Forecasting*, *16*(4), 521 - 530.

Billah, B.; King, M. L.; Snyder, R. D. and Koehler, A. B. (2006), 'Exponential smoothing model selection for forecasting', *International Journal of Forecasting*, *22*(2), 239 - 247.

Collopy, F. and Armstrong, J. (1992), 'Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations', *Management Science*,

38, 1392-1414.

Crone, S. and Kourentzes, N. (2011), 'Automatic Model Selection of Exponential Smoothing - an empirical evaluation of Trace Errors in forecasting for Logistics', *31st Annual International Symposium on Forecasting - ISF 2011*, June 26-29, 2011, Prague, Czech Republic.

Davydenko, A. and Fildes, R. (2013), 'Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts', *International Journal of Forecasting*, *29*(3), 510-522.

Fildes, R. (2001), 'Beyond forecasting competitions', *International Journal of Forecasting*, *17*, 556-560.

Fildes, R. (1989), 'Evaluation of aggregate and individual forecast method selection rules', *Management Science*, *39*, 1056-1065.

Fildes, R.; Hibon, M.; Makridakis, S. and Meade, N. (1998), 'Generalising about univariate forecasting methods: further empirical evidence', *International Journal of Forecasting*, *14*, 339-358.

Goodrich, R. L. (2000), 'The Forecast Pro methodology', *International Journal of Forecasting*, *16*(4), 533-535.

Hyndman, R. J. and Billah B. (2003), 'Unmasking the Theta method', *International Journal of Forecasting*, *19*(2), 287-290.

Hyndman, R. J. and Khandakar, Y. (2008), 'Automatic Time Series Forecasting: The forecast Package for R', *Journal of Statistical Software, 27*(3), 1-22.

Hyndman, R. J. and Athanasopoulos G. (2012). *Forecasting: principles and practice*, Melbourne, Australia: OTexts. http://otexts.com/fpp/.

Hyndman, R. J.; Koehler, A. B.; Snyder, R. D. and Grose, S. (2002), 'A state space framework for automatic forecasting using exponential smoothing methods', *International*

Journal of Forecasting, *18*(3), 439 - 454.

Makridakis S.; Wheelwright S.C. and Hyndman R.J. (1998), *Forecasting: Methods and Applications* (3rd ed.), New York: John Wiley and Sons.

Makridakis, S. and Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting*, *16*(4), 451 - 476.

Makridakis, S. and Winkler, R. L. (1989), 'Sampling distributions of post-sample forecasting errors', *Applied Statistics-Journal of the Royal Statistical Society Series C*, *38*, 331-342.

Meade, N. (2000), 'Evidence for the selection of forecasting methods', *Journal of Forecasting*, *19*, 515-535.

Ord, K. and Fildes, R. (2013), *Principles of Business Forecasting*, South-Western Cengage Learning.

Pant, P. N. and Starbuck, W. H. (1990), 'Innocents in the Forest - Forecasting and Research Methods', *Journal of Management*, *16*, 433-460.

Petropoulos, F., and Nikolopoulos, K. (2013), 'Optimizing Theta model for monthly data', In Proceedings: *5th International Conference on Agents and Artificial Intelligence – ICAART 2013*, February 15-18, 2013, Barcelona, Spain.

Shah, C. (1997), 'Model selection in univariate time series forecasting using discriminant analysis', *International Journal of Forecasting*, *13*, 489-500.

Appendices

A1. Deseasonalization of the raw data

Deseasonalization of the raw data is achieved by applying a multiplicative classical decomposition (Makridakis, Wheelwright and Hyndman, 1998). Monthly seasonal indices are stored. The core extrapolative forecasting methods are then applied to the deseasonalized

data. Finally, statistical forecasts derived from this procedure are reseasonalized by multiplying each point forecast with the corresponding seasonal index. The steps of this procedure are presenting in the following:

1. Deseasonalize the raw data using multiplicative classical decomposition.
2. Store the seasonal indices.
3. Apply statistical forecasting models on the deseasonalized data.
4. Reseasonalize the final forecast using the stored indices.

A2. The Theta model

The Theta model (Assimakopoulos and Nikolopoulos, 2000) proposed the decomposition of the data in two or more so-called "Theta lines". The decomposition itself takes place to a seasonally adjusted series and is based on the modification of the local curvature by using a dedicated coefficient (θ). Upon the selection of a unique θ coefficient, a Theta line can be calculated, maintaining the mean and the slope of the data. At the same time, the selection of appropriate coefficients enables the improvement of the short or the long-term behaviour of the series. Originally, just two Theta lines were used, with θ values 0 and 2. Theta line (0) is nothing more than a straight line, representing the series slope without any curvatures. This line is extrapolated by simple linear regression. Theta Line (2) describes a line with double the curvatures from the original data. This line is forecasted using Simple Exponential Smoothing. The forecasts derived from the two Theta Lines are then combined with equal weights. Finally, forecasts are reseasonalized. Hyndman and Billah (2003) proved that the classic Theta model, where only two Theta Lines (0 and 2) are used, is equivalent to Simple Exponential Smoothing plus a deterministic trend, equal to half the trend of the original series.

While more general forms of the Theta model have been considered (e.g. Petropoulos and Nikolopoulos, 2013), by introducing more Theta Lines or unequal weights for combining the final forecasts, the current research uses the classic form of the Theta model, as implemented by Hyndman and Billah (2003), using the thetaf() function of the *forecasting* package for *R* statistical software.

A3. Error measures

Let $y_t(i)$ be the actual value of series *i* for time period *t* and $\hat{y}_t^m(i|h)$ be the point forecast of the same series for method *m* at forecast origin *t* for lead time *h*, then $EM_t^m(i|h)$ is the error measure of series *i* for method *m* at origin *t* for lead time *h*. Error Measure (*EM*) may be one of the following:

- Signed Error (E): $E_t^m(i|h) = y_{t+h}(i) - \hat{y}_t^m(i|h)$
- Squared Error (SE): $SE_t^m(i|h) = E_t^m(i|h)^2$
- Absolute Error (AE): $AE_t^m(i|h) = |E_t^m(i|h)|$
- Absolute Percentage Error (APE):$APE_t^m(i|h) = \left|\frac{E_t^m(i|h)}{y_{t+h}(i)}\right| \cdot 100$ (%)

Let us define the error made in forecasting series from time origins t_1 to t_2 averaged over horizons h_1 to h_2 as:

$$Mean\ EM_{t_1,t_2}^m(i|h_1,h_2)_i = \frac{1}{h_2 - h_1 + 1}\sum_{h=h_1}^{h_2}\left(\frac{1}{t_2 - t_1 + 1}\sum_{t=t_1}^{t_2} EM_t^m(i|h)\right)$$

Mean Absolute Percentage Error (*MAPE*) is the Mean *APE* summarized across all *N* time series, as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} Mean\ APE_{t_1,t_2}^m(i|h_1,h_2)$$

that is the mean absolute percentage error averaged over series, forecast horizons and origins. The *MdAPE* values given in the text are the $Median(Mean\ APE^{m}_{t_1,t_2})$.

The relative Mean Absolute Error for a series *i* can be defined as:

$$r_i = \frac{MAE_i}{MAE_i^b}$$

where MAE_i^b – MAE for baseline forecast for series i, MAE_i^m – MAE for method *m* for series i. MAE_i^b and MAE_i^m can be obtained from the arithmetic mean absolute error averaged across all forecast origins and forecast horizons h_1 to h_2 for series *i*:

$$MAE_i^b = Mean\ AE^b_{t_1,t_2}(i|h_1,h_2)$$

$$MAE_i^m = Mean\ AE^m_{t_1,t_2}(i|h_1,h_2)$$

Davydenko and Fildes (2013) showed that when making comparisons between methods, the use of arithmetic means rather than geometric can lead to misinterpretations. Instead, they proposed the use of a geometric average relative MAE.

$$AvgRelMAE = \left(\prod\nolimits_{i=1}^{m} r_i\right)^{1/m}$$

As is standard practice, Absolute Percentage Errors and Squared Errors are also used in the simple selection approaches in order to select the most promising single forecasting approach from a specific pool of methods. The average *Past Forecast Performance* (*PFP*) of series *i* for a method *m* for origins t_1 through t_2 may be calculated as the performance over a fixed lead time (h) or multiple lead times (h_1 to h_2) measured by an *EM* as follows:

Single lead time: ${}_{EM}PFP^m_{t_1,t_2}(i|h) = \frac{1}{t_2-t_1}\sum_{t=t_1}^{t_2-h} EM_t^m(i|h)$

Multiple lead times: ${}_{EM}PFP^m_{t_1,t_2}(i|h_1,h_2) = \frac{1}{h_2-h_1+1}\sum_{h=h_1}^{h_2} {}_{EM}PFP^m_{t_1,t_2}(i|h)$

Note that in the special case of single horizon, $h_1=h_2$, the second equation is equivalent to the first one.

A4. Model Selection Rules

Assuming that model selection will be performed at the T_2+k origin, the *PFP* is measured between origins T_1 to T_2+k as is the fitted performance. The method to be selected is the one with the most promising performance. To this direction, the method with the minimum error (the smallest *PFP*), for the different lead times, is selected:

Single lead time: $Best\ Method = argmin\left[_{EM}PFP^{m}_{T_1,T_2+k}(i|h), m\right]$

Multiple lead times: $Best\ Method = argmin\left[_{EM}PFP^{m}_{T_1,T_2+k}(i|h_1,h_2), m\right]$

Argmin is a function (here of *m*, the alternative forecasting methods) which identifies the method whose past forecast performance for series *i* and horizon *h* (*PFP*(*i*|*h*)) is the minimum.

In the following, the four simple rules implemented and examined in this research are defined. These rules are applied, as previously mentioned, in a rolling origin matter. As such, the most appropriate method identified and applied for the calculation of the forecasts for the next origin may change over time. Nevertheless, in each origin *h* point forecasts are calculated. Note that in all cases *m* is the index referring to each one of the methods examined by a specific rule.

Rule 1. Use the method with best fit as measured by the minimum one-step ahead in sample Mean Squared Error:

$$Best\ Method_{T_{2+k}}\ (i|for\ all\ lead\ times) = argmin\left[_{SE}PFP^{m}_{T_1,T_2+k}(i|1), m\right]$$

Rule 2. Use the method with the best out-of-sample one-step-ahead forecast error, in terms of Mean Absolute Percentage Error, and apply that method to forecast for all lead times:

$$Best\ Method_{T_2+k}(i|for\ all\ lead\ times) = argmin\left[{}_{APE}PFP^{m}_{T_1,T_2+k}(i|1), m\right]$$

Rule 3. Use the method with best out-of-sample *h*-step-ahead forecast, in terms of Mean Absolute Percentage Error, and apply this method to forecast for just the same lead time:

$$Best\ Method_{T_2+k}(i|for\ lead\ time\ h) = argmin\left[{}_{APE}PFP^{m}_{T_1,T_2+k}(i|h), m\right]$$

Rule 4. Use the best out-of-sample 1-18-steps-ahead, in terms of Mean Percentage Absolute Error, method to forecast for all lead times:

$$Best\ Method_{T_2+k}(i|for\ all\ lead\ times) = argmin\left[{}_{APE}PFP^{m}_{T_1,T_2+k}(i|1{,}18), m\right]$$

Table 1. Forecasting methods included in the experiment

#	**Method**	**Short Name**	**Applied to**
1	Naive	Naive	Raw data
2	Naive 2	DNaive	Deseasonalized data
3	SES	Expsmoo	Raw data
4	SES 2	DExpsmoo	Deseasonalized data
5	Holt	Holt	Raw data
6	Holt 2	DHolt	Deseasonalized data
7	Holt-Winters	HoltWint	Raw data
8	Damped	Damp	Raw data
9	Damped 2	DDamp	Deseasonalized data
10	Damped with multiplicative seasonality	DampMult	Raw data
11	Theta	Theta	Deseasonalized data
12	ARIMA	ARIMA	Raw data

Table 2. Segmenting the data set: number of time series per segment

Segment	**Number of series**
Entire data set	998
Predictable	694
Unpredictable	304
Trended	894
Non-trended	104

Seasonal	608
Non-seasonal	390
Dominant Method	428
Non-dominant method	570
Stable performance	249
Unstable performance	749

Table 3. Forecasting accuracy measured by *MdAPE* (%) averaged across all lead times and by series segments for the test data

	Entire data set	Predictable	Unpredictable	Trended	Non-trended	Seasonal	Non-seasonal	Dominant	Non-	Stable	Unstable
Naive	9.8	10.1	9.6	9.1	20.2	13.1	4.4	11.6	8.3	9.8	9.8
DNaive	7.9	7.3	9.3	7.3	15.7	9.6	4.6	9.2	7.2	6.0	9.1
Expsmoo	9.0	8.9	9.1	8.2	18.4	12.0	4.2	10.6	7.7	8.8	9.1
DExpsmoo	7.3	6.7	9.0	6.8	15.1	8.8	4.3	8.0	6.8	5.6	8.3
Holt	9.8	9.5	10.1	8.6	20.1	13.3	5.0	11.7	8.1	9.5	9.8
DHolt	8.2	7.2	10.7	7.6	16.0	10.8	5.4	8.8	7.5	5.8	9.5
HoltWint	8.5	7.6	10.8	8.0	17.1	10.8	5.5	9.5	7.9	6.0	9.5
Damp	9.1	8.8	9.6	8.3	18.7	11.9	4.3	10.9	7.9	8.8	9.1
DDamp	7.5	6.6	9.5	6.9	15.1	9.2	4.5	8.2	6.8	5.4	8.7
DampMult	7.5	6.5	9.3	7.1	15.6	9.2	4.7	7.8	7.1	5.4	8.7
Theta	7.4	6.5	9.0	6.8	14.7	9.2	4.5	7.8	6.6	5.1	8.3
ARIMA	7.7	7.1	9.0	7.2	15.6	9.7	4.8	8.4	7.0	5.4	8.5

Perfect Information	5.3	4.7	6.3	4.8	10.5	6.7	2.7	5.9	4.8	3.6	5.9

Table 4. Percentage (%) of cases where individual selection leads to improved forecasting accuracy when compared to (i) aggregate selection and (ii) combination.

Methods in selection pool	Average Correlation	Number of cases	% of cases Individual Selection performed better							
			vs. (i) Aggregate				vs. (ii) Combination			
			Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
2-4	Low	611	27.8	**67.1**	48.4	**88.1**	33.4	**75.0**	**83.1**	**90.0**
	High	170	21.2	**72.4**	**61.2**	**79.4**	18.8	**60.6**	**70.6**	**80.6**
5-8	Low	2712	20.7	**57.4**	**55.0**	**96.1**	38.6	**80.7**	**89.5**	**97.5**
	High	291	8.9	**82.8**	**67.0**	**94.8**	9.6	**84.5**	**82.1**	**95.9**
9-12	Low	295	6.4	49.5	44.7	**100.0**	45.1	**91.2**	**95.3**	**100.0**
	High	4	0.0	**100.0**	25.0	**100.0**	25.0	**100.0**	**100.0**	**100.0**

Table 5. Percentage (%) of cases for each segment where individual selection leads to improved forecasting accuracy when compared to (i) aggregate selection and (ii) combination.

Segment	Average Correlation	Number of cases	% of cases Individual Selection performed better							
			vs. (i) Aggregate				vs.(ii) Combination			
			Rule 1	Rule 2	Rule 3	Rule 4	Rule 1	Rule 2	Rule 3	Rule 4
Predictable	Low	3945	25.8	43.7	32.8	**51.0**	41.1	**77.0**	**75.8**	**87.8**
	High	138	26.8	**59.4**	26.1	36.2	22.5	47.1	32.6	38.4
Unpredictable	Low	51	**76.5**	**74.5**	**86.3**	**90.2**	11.8	17.6	**51.0**	**56.9**
	High	4032	**55.1**	**76.9**	**92.3**	**89.2**	3.2	23.7	45.9	**52.1**
Trended	Low	464	43.5	**65.1**	**64.0**	**81.0**	14.2	**72.8**	**66.6**	**85.8**
	High	3619	22.4	48.5	46.0	**72.7**	16.6	**71.3**	**71.3**	**84.6**
Non-Trended	Low	3964	35.3	45.4	**76.6**	**67.8**	35.7	**68.8**	**85.3**	**83.4**
	High	119	**52.9**	42.9	44.5	44.5	6.7	26.1	31.9	37.8
Seasonal	Low	3870	27.1	49.8	**83.7**	**90.9**	40.5	**74.0**	**98.7**	**99.1**
	High	213	23.0	**52.6**	**91.1**	**93.4**	10.8	31.9	**82.2**	**83.1**
Non-Seasonal	Low	3	33.3	0.0	33.3	33.3	0.0	33.3	0.0	0.0
	High	4080	17.5	19.7	20.4	21.7	1.5	6.3	2.3	4.4
Dominant method	Low	4002	27.2	36.4	31.7	47.9	**55.4**	**70.0**	**81.2**	**87.7**
	High	81	27.2	38.3	33.3	32.1	18.5	46.9	44.4	**53.1**
Non-dominant method	Low	953	**58.3**	**82.7**	**78.5**	**87.1**	30.5	**81.0**	**73.9**	**87.7**
	High	3130	47.5	**83.0**	**77.2**	**81.1**	11.6	**64.6**	47.7	**68.6**
Stable	Low	4009	**59.2**	**55.8**	**55.0**	**71.1**	**73.9**	**78.5**	**82.7**	**84.1**
	High	74	35.1	32.4	**56.8**	**58.1**	13.5	20.3	40.5	41.9
Unstable	Low	115	**78.3**	**80.9**	**84.3**	**96.5**	3.5	5.2	23.5	45.2
	High	3968	32.8	**51.9**	32.8	**70.8**	2.4	21.8	13.7	**50.1**

Table 6. Best selection approach for each segment

Segment	Best Practice
Entire Data Set	Individual selection (Rule 4) using a high number of low correlated methods
Predictable	Combination using a medium number of high correlated methods
Unpredictable	Individual selection (Rule 4) using a high number of high correlated methods
Trended	Individual selection (Rule 4) using a high number of high correlated methods
Non-Trended	Combination using a medium number of high correlated methods
Seasonal	Individual selection (Rule 3) using a high number of low correlated methods
Non-Seasonal	Aggregate selection using a high number of high correlated methods
Dominant method	Aggregate selection using a medium number of high correlated methods
Non-dominant method	Individual selection (Rule 4) using a high number of high correlated methods
Stable	Individual selection (Rule 4) using a medium number of high correlated methods
Unstable	Individual selection (Rule 4) using a high number of high correlated methods

Table 7. *MdAPE*s (%) for aggregate selection, simple combination, individual selection and best practice, analysed by segments.

Segment	Aggregate Selection	Combination	Individual Selection (Rule 4)	Performance of Best Practice	DDamp (Benchmark)	Improvement
Entire Data Set	7.4	7.6	7.2	7.1	7.5	4.7%
Predictable	6.5	6.9	6.6	6.5	6.6	2.4%
Unpredictable	9.6	8.8	8.7	8.5	9.5	11.0%
Trended	6.8	7.0	6.8	6.7	6.9	2.2%
Non-Trended	15.6	15.8	15.2	14.8	15.1	2.0%
Seasonal	9.2	9.5	9.0	8.9	9.2	3.2%
Non-Seasonal	4.3	4.4	4.6	4.2	4.5	5.4%
Dominant Method	7.9	8.4	7.9	7.8	8.2	5.2%
Non-dominant method	7.1	6.9	6.8	6.7	6.8	1.7%
Stable	5.4	5.9	5.4	5.1	5.4	5.3%
Unstable	8.4	8.3	8.3	8.3	8.7	4.8%

Table 8. Out-of-sample performance (*MdAPE*s) of best practice for the case study.

Segment	Number of series	Aggregate Selection	Combination	Individual Selection (Rule 4)	Performance of Best Practice	DDamp (Benchmark)	Improvement
Entire Data Set	251	31.4	27.8	29.3	28.6	30.1	5.0%
Predictable	142	28.8	26.0	26.3	24.8	24.3	-1.9%

Unpredictable	109	37.5	32.7	35.4	34.1	38.2	10.8%
Trended	126	30.7	27.4	28.5	28.4	28.6	0.8%
Non-Trended	125	34.0	28.3	30.8	26.9	32.8	18.0%
Seasonal	48	24.3	21.8	20.5	21.4	20.0	-6.7%
Non-Seasonal	203	35.2	30.6	32.8	35.2	34.4	-2.4%
Dominant Method	122	35.2	32.0	34.6	35.2	37.4	5.8%
Non-dominant method	129	24.6	24.9	24.3	24.2	24.4	0.9%
Stable	63	27.4	26.2	26.4	26.4	25.9	-2.0%
Unstable	188	33.1	28.3	30.8	31.8	31.7	-0.5%